이상지질혈증 연관 인자를 활용한 얕은 학습 기반 HDL-콜레스테롤 예측

양수빈 [1], 김민태 [1], 권수빈 [1], 김학재 [2], 정태경 [3], 이성주 [1†]
상명대학교 [1], ㈜클래스액트 [2]
{201921007[1], 201820985[1], 202020999[1], peacfeel[1†]}@smu.ac.kr
{krunvis}@gmail.com,
{ttjeong}@hallym.ac.kr

HDL Cholesterol Prediction based on Shallow Learning using Dyslipidemia-related factors

Subhin Yang[1], Mintae Kim[1], Subin Kwon[1], Hakjae Kim[2], Taikyeong Jeong[3], Sungju Lee[1†]
Department of Software, Sangmyung University[1]
CLASSACT Incorporated[2]
School of Artificial intelligence Convergence Hallym University[3]

요 약

이상지질혈증의 연관 인자를 파악하고 발병에 대한 조기 진단 및 관리하는 것은 중요한 문제이다. 본 논문에서는 연관 인자 특징들을 이용하여 기계학습 기반 HDL-콜레스테롤을 예측하는 방법을 제안한다. 제안 방법은 기계 학습에 이용한 특징의 개수가 최대 12개로 많지 않기 때문에 얕은 학습(Shallow Learning)기반의 MLP(Multi Layer Perceptron)을 이용한다. 또한 정확도를 개선하기 위해서 각 특징에 대한 HDL-콜레스테롤의 고위험군과 저위험군의 유사성을 분석하여 중요한 특징을 선택하는 방법을 적용한다. 실험결과, 모든 연관 인자 특징들, HDL-콜레스테롤과 연관이 있다고 알려져 있는 특징들, 그리고 중요도를 파악하여 선택한 특징들의 MLP을 이용하여 예측한 정확도는 각각 65.3%, 67.4%, 그리고 70.8%로 정확도를 최대 5.5% 개선할 수 있음을 확인한다.

1. 서 론

사회가 발달함에 따라 만성질환인 비만과 이상지질혈증 유병률이 점차적으로 증가하고 있으며[1], HDL-콜레스테롤과의 연관 인자를 분석하여 이상지질혈증 발병에 대한 조기 진단 및 관리는 중요한 문제이다. 이상지질혈증의 유병 원인으로 중년기 이후의 성인에게서 하지불안증후군 등에 의한 수면의 질 저하[2], 비만, 성별 등이 이상지질혈증 발생 위험 요인으로 알려져 있다[3, 4]. 특히 수면의 질 낮을수록 이상지질혈증의 위험인자 중 하나인 HDL-콜레스테롤이 감소한다는 연구가 보고되었다[3]. 이상지질혈증의 조기 진단 및 관리문제를 해결하기 위해서는 수면의 질과 비만 정도와 같은 연관인자등을 활용하여 이상지질혈증 발병에 대한 예측 방법이 필요하다.

본 논문에서는 혈액 수집이 요구되지 않으면서 이상지질혈증과 관련이 있을 수 있는 12가지 연관 인자 특징들을 이용하여 HDL-콜레스테롤을 예측할 수 있는 모델을 설계한다. 또한, 이상지질혈증과 관련된 연관 인자는 성별, 나이, 체지방률, 그리고 수면의 질을 이용한다. 특히 수면의 질의 평가를 위해서 신뢰도와 타당성이 입증된 Pittsburgh Sleep Quality Index(PSQI)[1, 2, 3, 5]을 이용한다. 제안 방법은 기계 학습에 이용한 특징의 개수가 최대 12개로 많지 않기 때문에 얕은 학습(Shallow Learning)기반의 MLP(Multi Layer Perceptron)을 이용한다. 최근에도 RNN을 이용한 제 2형 당뇨병 예측모델[6]과 같은 질병 예측 모델이 제안되었고, 이와 유사한 방법을 사용하여 혈액 수집 특징을 제외한 신체 계측치와 PSQI을 학습시킨 MLP을 이용해 HDL을 예측하는 모델을 제안한다. 또한 정확도를 개선하기 위해서 각 특징에 대한 HDL-콜레스테롤의 고위험군과 저위험군의 유사성을 분석하여 중요한 특징을 선택하는 방법을 적용한다. 이전 연구에서는 이상지질혈증 환자와 비환자의 관계를 MSE(Mean Squared Error)로 분석해 특징 분포도가 유사하지 않은 특징을 중요 특징으로 선택하는 방법[7]과 유사한 방법을 사용한다. 이전 연구와는 혈액 수집 여부, 특징 분포도 분석 방식, 그리고 예측 값의 차이가 있으며, 중요한 임상 데이터 특징을 선택하기 위해 HDL-콜레스테롤 범위 별 특징의 분포를 비교하여 중요한 특징을 선택한다. 즉, MLP에 이용하는 중요한 임상 데이터는 HDL-콜레스테롤 낮은 군과 보통 군, 그리고 높은 군의 특징 분포 범위를 비교하여 우선순위가 높은 특징 선택의 결과를 바탕으로, 총 10가지 특징과 8가지의 특징을 선택한다.

실험결과, 모든 연관 인자 특징들(Features Set-A), HDL-콜레스테롤과 연관이 있다고 알려져

있는 특징들(Features Set-B), 그리고 중요도를 파악하여 선택한 특징들(Features Set-C)의 MLP을 이용하여 예측한 정확도는 각각 65.3%, 67.4%, 그리고 70.8로 정확도를 최대 5.5% 개선할 수 있음을 확인한다.

본 논문의 구성은 다음과 같다. 2장에서는 연관 인자 특징, 임상적 정상범위, 중요도가 높은 특징을 선택하는 방법을 설명하며, 3장에서는 실험 환경과 실험 결과를 설명한다. 그리고 4장에서는 결론을 설명한다.

2. 본 론

2.1. 연관 인자

본 논문에서는 HDL-콜레스테롤을 예측하기 위한 임상 데이터를 연세 의료원에서 제공하는 건강 관리 임상시험 데이터를 사용한다. 180명의 8주간 1차, 2차로 방문으로 총 360개 (20~63세의 남녀, 남성: 125명, 여성, 55명)의 샘플로 진행한다. 연관 인자 특징 개수는 총 12가지(AGE, SEX, BMI, PSQI, Muscle, Fat, SBP, DBP, HR, Waist, Fat Percentage, WHR)이며 혈액 수집을 요구하지 않는 특징이다.

2.2. 임상적 HDL-콜레스테롤의 정상 범위

이상지질혈증은 4가지의 혈중 지질(Total Cholesterol, LDL, TG, HDL)의 농도를 통해 임상적 정상범위로 이상지질혈증 진단이 가능하다[7, 8]. 표 2은 이상지질혈증 기준 중 하나인 HDL의 임상적 정상범위를 보여준다.

표 1. HDL의 임상적 정상/위험 범위

	진단
40 미만	낮음 (low)
60 이상	높음 (high)

2.3. 중요한 특징 선택

혈액 수집을 요구하지 않는 12가지의 특징 중에서, 각 특징 따른 HDL-콜레스테롤 분포를 낮은 군(HDL<40), 보통 군(40 ≤ HDL<60), 그리고 높은 군(HDL≥60)을 각각 비교하여 중요한 특징을 선택한다. 즉, 각 특징 분포도가 유사하지 않으면 HDL-콜레스테롤을 예측할 수 있는 특징 변화가 있다고 간주하여 정확도를 개선할 수 있는 중요 특징으로 선택한다. 그림 1은 12가지의 특징 중, 낮은 군의 특징 분포도와 보통 군, 높은 군의 특징 분포도가 유사하지 않다고 간주되는 예(AGE)를 보여준다. 또한, 그림 2는 특징의 HDL-콜레스테롤 분포도가 유사한 예(FatPercentage)의 분포도를 보여준다. 이러한 특징 분포도가 유사하지 않은 것을 우선순위로 정하며, 총 12가지의 특징 중에서 하위 순위 두 가지의 특징(FatPercentage, WHR)을 제외시킨 Features Set-C-1과 하위 순위 네 가지의 특징(FatPercentage, WHR, PSQI, HR)을 제외시킨 Features Set-C-2를 구성한다. 즉, 제외시킨 두 가지의 특징과 네 가지의 특징은 HDL을 예측하는 중요한 특징으로 선택하지 않는다(그림 2 참고). 반면에 제외되지 않은 나머지의 특징은 세 개의 군이 특징 분포도가 유사하지 않아 정확도를 개선할 수 있는 중요 특징으로 선택한다(그림 1 참고).

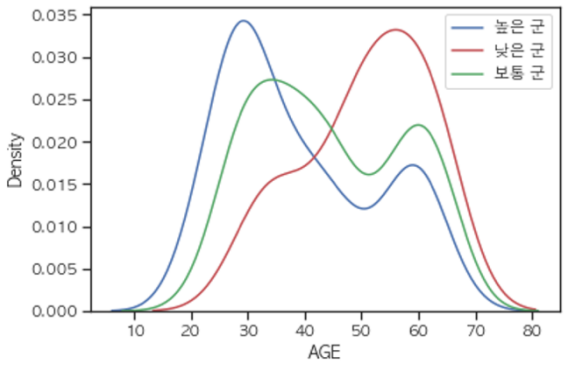

그림 1. 특징 분포도가 유사하지 않은 특징(AGE)의 분포도

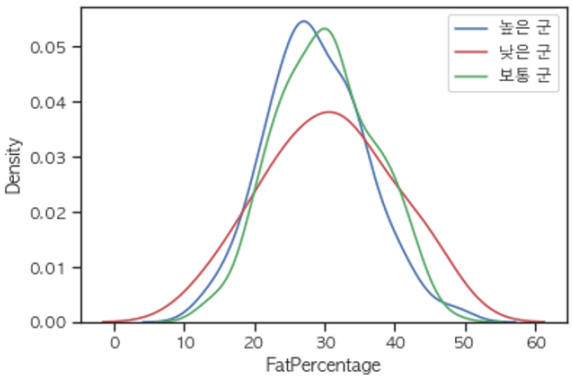

그림 2. 특징 분포도가 유사한 특징(FatPercentage)의 분포도

3. 실험 환경 및 실험 결과

3.1. 실험 환경

본 연구의 데이터는 180명의 8주간 방문을 합친 총 360개 (20~63세의 남녀, 남성: 125명, 여성, 55명)의 임상 데이터로 진행하였다. MLP의 학습데이터는 288개의 훈련 집합 및 72개의 테스트 집합으로 구성하였다.

표 2는 모든 연관 인자 특징들(Features Set-A), HDL-콜레스테롤과 연관이 있다고 알려져 있는 특징들(Features Set-B), 그리고 중요도를 파악하여 선택한 특징들(Features Set-C)의 각 특징 구성 요소를 보여준다.

표 2. 다양한 Feature Set

	연관 인자
FS-A	AGE, SEX, BMI, PSQI, Muscle, Fat, SBP, DBP, HR, Waist, Fat Percentage, WHR

FS-B	AGE, SEX, BMI, PSQI
FS-C-1	AGE, SEX, BMI, PSQI, Muscle, Fat, SBP, DBP, HR, Waist
FS-C-2	AGE, SEX, BMI, Muscle, Fat, SBP, DBP, Waist

본 논문에서는 HDL 예측을 위해 다층 퍼셉트론(MLP; Multi-Layer Perceptron)을 활용하여 모형을 구성하고, 예측 정확도를 측정하였다. 임상 데이터에서 추출한 Features Set-A, Features Set-B, Features Set-C-1, 그리고 Features Set-C-2를 Standard Scaler을 사용해 데이터 정규화를 진행하였다. 데이터를 80% 비율인 288개의 훈련집합으로 나눠 입력층에 넣어 32가지의 노드를 가진 2개의 은닉층을 거친 뒤, 출력층 노드를 통해 HDL 예측값을 산출하는 과정을 배치 사이즈(batch_size)를 4로 적용하였고 학습 횟수(epoch)는 150이었다.

3.2. 실험 결과

표 3은 FS-A, FS-B, FS-C-1, 그리고 FS-C-2에 대한 MLP의 예측 정확도를 보여준다. 허용 오차(즉 손실 오차)를 ±10으로 설정했을 때, 예측 정확도는 각각 65.3%, 67.4%, 70.8%, 그리고 63.9%로 측정되었다. HDL-콜레스테롤과 연관 인자로 알려진 FS-B는 FS-A(모든 특징) 보다 2.1% 높은 성능을 제공하였음을 확인하였다. 또한 FS-C-1(두 개의 특징을 제거)는 FS-A보다 5.5% 높은 성능을 제공하였으며, FS-C-2(네 개의 특징을 제거) 보다 6.9% 높은 정확도를 제공하였음을 확인하였다.

표 3. 특징에 따른 MLP의 HDL 예측 정확도

허용 오차	특징에 따른 정확도 (%)			
	FS-A	FS-B	FS-C-1	FS-C-2
± 10	65.3	67.4	70.8	63.9

4. 결 론

본 논문에서는 혈액 수집을 요구하지 않는 특징으로 신체계측치와 PSQI 로 구성된 12 가지 연관 인자 특징을 이용하고 얕은 학습 방법으로 알려진 MLP 을 이용하여 HDL 을 예측하는 방법을 제안하였다. 또한 기계 학습의 정확도를 개선하기 위해서, 특징 분포도가 유사하지 않으면 HDL-콜레스테롤을 예측할 수 있는 특징 변화가 있다고 간주하여 정확도를 개선할 수 있는 중요한 특징으로 선택하였다. 실험결과, 모든 연관 인자 특징들(FS-A), HDL-콜레스테롤과 연관이 있다고 알려져 있는 특징들(FS-B), 그리고 중요도를 파악하여 선택한 특징들(FS-C)의 MLP 을 이용하여 예측한 정확도는 각각 65.3%, 67.4%, 그리고 70.8%로 정확도를 최대 5.5% 개선할 수 있음을 확인한다.

ACKNOWLEDGEMENT

본 연구는 2020 년도 중소벤처기업부의 기술개발사업 지원에 의한 연구임 [S2935743]

참 고 문 헌

[1] 한아름, 수면과 대사증후군과의 관계, 연세대학교 대학원 석사학위논문, 2008.

[2] Y. G. Bak, H. S. Park, Quality of Sleep and Serum Lipid Profile in Patients with Restless Legs Syndrome, Journal of Korean Academy of Nursing, 41, 3, 344-353, 2011.

[3] E. J. Lee, S. G. Kang, J. H. Shin, S. W. Song, Y. N. Hwang, K. S. Ryu, Relationship between Sleep Quality and Metabolic Syndrome and Inflammatory Markers in Middle-aged Men in Korea, Korean Journal of Family Medicine, 30, 5, 344-351, 2009.

[4] M. Y. Jeon, W. H. Choi, Y. M. Seo, Risk Factors of Dyslipidemia and Related Factors of Medication Adherence in Korea Adults: KNHANES 2013-2015, 기초간호자연과학회지, 19, 3, 131-140, 2017.

[5] S. H. Shin, S. H. Kim, The realiability and validity testing of Korean version of the pittburgh sleep quality index, Journal of Convergence for Information Technology, 10, 11, 148-155, 2020.

[6] J. S. Jang, M. J. Lee, T. R. Lee, Development of T2DM Prediction Model Using RNN. Journal of Digital Convergence, 17, 8, 249-255, 2019.

[7] Seonmin Lee, Mintae Kim, Subin Yang, Hakjae Kim, Taikyeong Jeong, Sungju Lee, Important Clinical Data Selection for Machine Learning Accuracy of Hyperlipidemia Digonosis, Proceedings of Symposium of the Korean Institute of communications and Information Sciences, 1426-1427, 2021.

[8] I. Jeong, 고지혈증/비만. 대한외과학회 학술대회 초록집, 161-162, 2018.